THE ROYAL STATISTICAL SOCIETY

2006 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

PAPER III
STATISTICAL APPLICATIONS AND PRACTICE

The Society provides these solutions to assist candidates preparing for the examinations in future years and for the information of any other persons using the examinations.

The solutions should NOT be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids.

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of these solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of these solutions.

Note. In accordance with the convention used in the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Higher Certificate, Paper III, 2006. Question 1

(i) Null hypothesis: there is no difference between the population mean failure stresses of the 200°C and 250°C tempered steel. Alternative hypothesis: there is a difference between the mean failure stresses, that for 250°C being higher.

$n_1 = 8,\ n_2 = 7;\ \bar{x}_1 = 59.6,\ \bar{x}_2 = 63.6;\ s_1^2 = 17.4^2,\ s_2^2 = 20.1^2.$

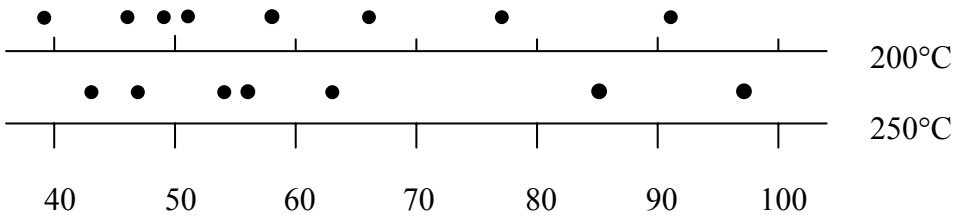
The pooled estimate of the assumed common variance (see part (ii)) is $s^2 = \left((7 \times 17.4^2) + (6 \times 20.1^2)\right)/13 = 349.491$ (so $s = 18.69$), with 13 d.f.

Thus the test statistic for testing that the population means are the same is

$$\frac{\bar{x}_2 - \bar{x}_1(-0)}{s\sqrt{\frac{1}{8} + \frac{1}{7}}} = \frac{4.0}{9.681} = 0.41,$$

which is referred to t_{13}. This is not significant at the 5% level (upper single-tailed 5% point is 1.771), so there is no evidence to reject the null hypothesis – it seems that the population means are the same.

(ii) The two underlying populations are assumed Normally distributed, with the same variance. A dot plot helps to check these assumptions:



Sample sizes are small but the ranges are similar and it may be reasonable to assume similar variances. However, Normality is in some doubt, as there is no central clustering and some skewness and/or outliers may be present.

Solution continued on next page

(iii) The Wilcoxon rank sum test (or, equivalently, the Mann Whitney U form of this test) is suitable. The null hypothesis is that the population median at 250°C is equal to that at 200°C, and the alternative is that it is greater. We first rank all 15 data items, as follows.

Data	39	43	46	47	49	51	54	56	58	63	66	77	85	91	97
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	A	*B*	*A*	*B*	*A*	*A*	*B*	*B*	*A*	*B*	*A*	*A*	*B*	*A*	*B*

A refers to 200°C, *B* to 250°C.

The rank sum for the smaller sample (*B*) is 2 + 4 + 7 + 8 + 10 + 13 + 15 = 59.

The required test is one-sided. For a 5% test, we refer this to the lower 5% point for the $W_{7,8}$ distribution as shown in the Society's statistical tables for use in examinations. This is 41 so, at the 5% level of significance, we cannot reject the null hypothesis; it appears that the two populations are the same in this regard.

(iv) Neither test supports the hypothesis of an increase. The nonparametric test is more suitable for these data because there is doubt regarding the assumption of Normality that underlies the *t* test.

Higher Certificate, Paper III, 2006. Question 2

(i) The analysis of variance table is as follows. Entries in *italics* are given in the question. The others need to be calculated.

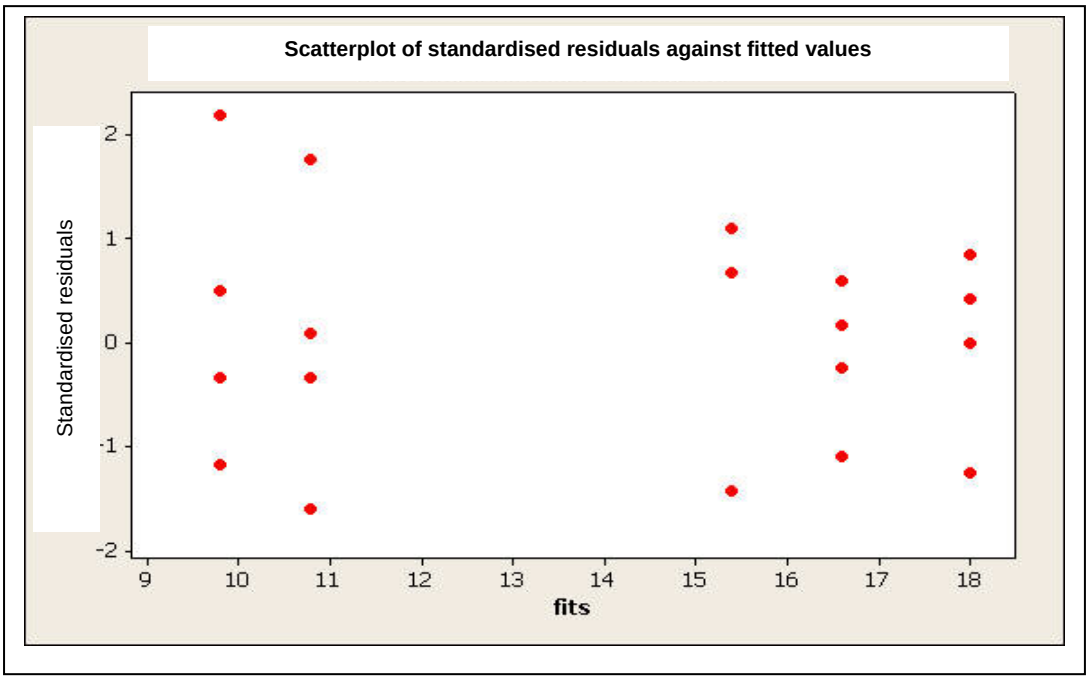
SOURCE	DF	SS	MS	*F* value
Percentages	4	*262.64*	*65.66*	9.25 Compare $F_{4,20}$
Residual	20	142.00	7.10	$= \hat{\sigma}^2$
TOTAL	24	*404.64*		

Upper critical points of $F_{4,20}$ are as follows:

5%	1%	0.1%
2.87	4.43	7.10

The *F* value for percentages is very highly significant; we have very strong evidence that not all the percentages of cotton are the same in terms of mean tensile strength of the synthetic fibre.

(ii) The fitted values are simply the sample means for the different percentages of cotton: (15%) 9.8; (20%) 16.6; (25%) 18.0; (30%) 15.4; (35%) 10.8. The graph suggests that where the mean is lower the variance tends to be slightly higher. The analysis is based on a model in which the residual term has constant variance, but apparent departures from this are not great and there is no reason to doubt the results seriously.



Note. Some points on this graph represent two coincident values.

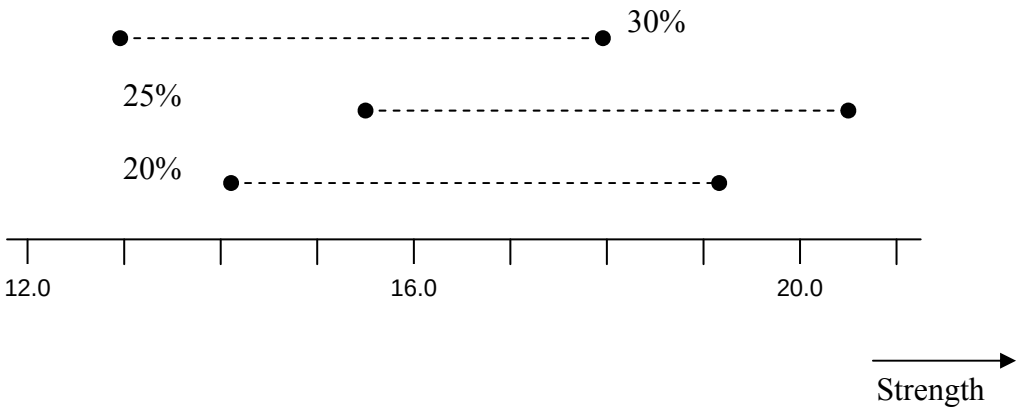
Note also the "false origin" on the fitted values axis.

Solution continued on next page

(iii) A 95% confidence interval for an individual mean is given by $\bar{x} \pm 2.086\hat{\sigma}/\sqrt{5}$ where 2.086 is the double-tailed 5% point of t_{20}.

Further, $\hat{\sigma}^2 = 7.10$, the residual mean square in the above analysis of variance. Thus the interval for 20% cotton is given by $16.6 \pm 2.086\sqrt{7.10}/\sqrt{5}$, i.e. it is (14.1, 19.1).

Similarly, for 25% the interval is (15.5, 20.5) and for 30% it is (12.9, 17.9).



On this evidence, 25% cotton should be used.

(iv) It is worth exploring just below 25%, and perhaps just above. Depending on how much work can be done, it may be possible to search for a maximum in this region.

Higher Certificate, Paper III, 2006. Question 3

(i) In a simple random sample (from a finite population) every possible selection of a sample of given size has the same probability of being chosen. This also has the effect that every individual in the population has the same probability of being selected for the sample.

(ii) Given a large population, of N items listed in some order (e.g. alphabetical) which is not related to trends in the characteristics being observed or measured, a systematic sample will be a valid alternative to simple random sampling. [If n items are required for the sample, with $N = nk$, take a random starting point among the first k in the list and every kth thereafter.]

(iii)(a) 100 claim to be regular users, so $\hat{p} = 0.5$.

(iii)(b) We have a contingency table as follows. The null hypothesis is that there is no association between faculty and library use. The expected frequencies are shown in brackets in each cell (e.g. $25.0 = 50 \times 100/200$).

		Library use		
		Regular	*Non-regular*	Total
Faculty	*Engineering*	25 (25.0)	25 (25.0)	50
	Business	42 (35.0)	28 (35.0)	70
	Arts	21 (17.5)	14 (17.5)	35
	Informatics	12 (22.5)	33 (22.5)	45
	Total	100	100	200

The value of the test statistic is

$$X^2 = \frac{(25-25.0)^2}{25.0} + \frac{(25-25.0)^2}{25.0} + \frac{(42-35.0)^2}{35.0} + \ldots + \frac{(33-22.5)^2}{22.5} = 14.00.$$

This is referred to χ^2_3. It is highly significant (the 1% critical point is 11.345). There is strong evidence of an association between faculty and library use.

Solution continued on next page

(iii)(c) $p_f - p_m$ is estimated by $\hat{p}_f - \hat{p}_m = \frac{35}{65} - \frac{65}{135} = 0.538 - 0.481 = 0.057$. The estimated variance of $\hat{p}_f - \hat{p}_m$ is given by

$$\frac{\hat{p}_f\left(1-\hat{p}_f\right)}{n_f} + \frac{\hat{p}_m\left(1-\hat{p}_m\right)}{n_m} = 0.003823 + 0.001849 = 0.005673.$$

Thus the approximate 95% confidence interval for $p_f - p_m$ is given by 0.057 ± (1.96×√0.005673), i.e. it is (–0.09, 0.21).

This interval contains 0, so there is no real evidence of a difference in library use between the sexes.

(iii)(d) Several factors make the sample design unlikely to represent the student population closely. Faculties probably contain different numbers of students, there will be different male/female ratios in different faculties (though here there does not seem to be a sex difference in the results), and proportional sampling from these different sub-groups would be desirable.

The timing of the survey will be important, as timetables in different faculties, and other student activities, could affect the structure of a sample and bias it towards particular groups.

Students may be accompanied by others of similar interests or habits, and therefore sample units may not be independent.

Higher Certificate, Paper III, 2006. Question 4

(i)(a) Remaining *MA* values are: at 13, $\frac{1}{5}(13+14+13+14+15)=13.8$; at 14, $\frac{1}{5}(14+13+14+15+16)=14.4$; at 15, 15.0; at 16, 15.8; at 17, 16.2; and at 18, 16.4.

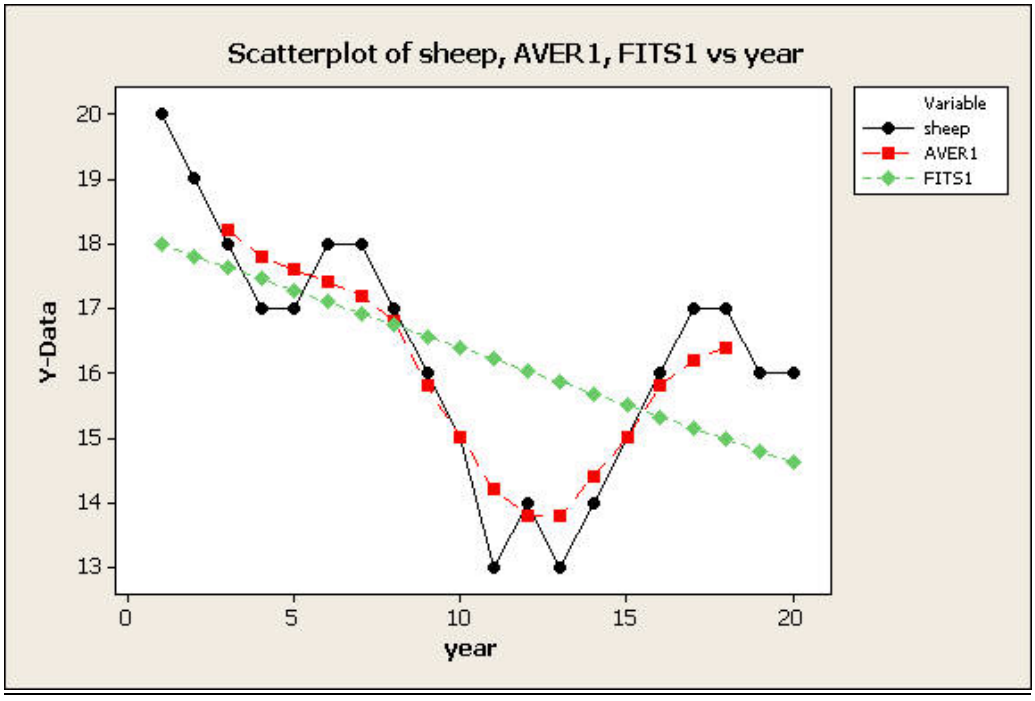
(i)(b) $$\text{Slope} = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{20}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{20}} = \frac{3305 - \frac{210\times 326}{20}}{2870 - \frac{210^2}{20}} = \frac{-118}{665} = -0.1774.$$

Intercept $= \bar{y} - \hat{b}\bar{x} = 16.3 - ((-0.1774)\times 10.5) = 18.163$.

Thus the regression equation is $sheep = 18.163 - 0.177\,year$.

(ii) The graph shows that the 5-point moving average (AVER1) best reflects the sharp changes in sheep population in years 10 to 15 while the regression (FITS1) line reflects the overall downward trend.

An unweighted MA is appropriate in the absence of periodic or cyclic effects.



(iii) The residual (error) terms have zero mean, constant variance and are uncorrelated. An assumption of Normality (so that the uncorrelatedness would imply independence) would also be needed it formal tests were to be carried out rather than only estimating the parameters. There does appear to be some sort of serial pattern in the data, rather than purely random variation from year to year. This makes the assumption of uncorrelatedness doubtful.

Higher Certificate, Paper III, 2006. Question 5

(i) The survivor function is $P(T > t) = \int_t^\infty \lambda^2 \theta e^{-\lambda\theta} d\theta$

$= \lambda^2 \left\{ \left[\theta\left(-\frac{1}{\lambda} e^{-\lambda\theta} \right) \right]_t^\infty + \frac{1}{\lambda} \int_t^\infty e^{-\lambda\theta} d\theta \right\} = \lambda t e^{-\lambda t} + \lambda \left[\frac{e^{-\lambda\theta}}{-\lambda} \right]_t^\infty = \lambda t e^{-\lambda t} + e^{-\lambda t}$,

as required.

[Alternatively, as we are only asked to show that the given function $S(t)$ is the survivor function, $\frac{dS}{dt} = -\lambda(1+\lambda t)e^{-\lambda t} + \lambda e^{-\lambda t} = -\lambda^2 t e^{-\lambda t}$, and $f(t)$ is therefore $-S'(t)$ as required.]

(ii) $L = \prod_{i=1}^n \lambda^2 t_i e^{-\lambda t_i}$, and hence $\log L = 2n \log \lambda + \sum_{i=1}^n \log t_i - \lambda \sum_{i=1}^n t_i$.

$\therefore \frac{d \log L}{d\lambda} = \frac{2n}{\lambda} - \sum t_i$ which on setting equal to zero gives that the maximum likelihood estimate is $\hat{\lambda} = \frac{2n}{\Sigma t_i}$. [Consideration of $\frac{d^2 \log L}{d\lambda^2}$ confirms that this is a maximum: $\frac{d^2 \log L}{d\lambda^2} = \frac{-2n}{\lambda^2} < 0$.]

So for the given sample, the value of $\hat{\lambda}$ is 20/707 = 0.0283.

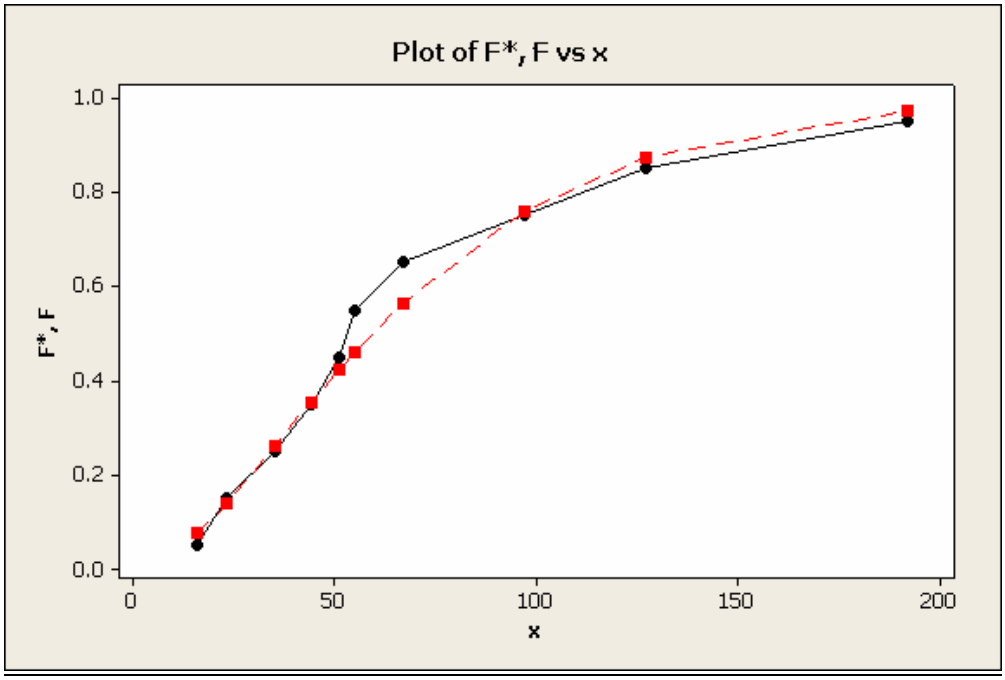
(iii) The estimated value of $S(240)$ is $(1 + (0.0283 \times 240))e^{-0.0283 \times 240} = 7.792e^{-6.792} = 0.00875$.

(iv) $F^*(x)$ is sometimes referred to as the "empirical cdf". Its values for this set of data are 1/20, 3/20, …, 17/20, 19/20 at x = 16, 23, …, 127, 192. Strictly speaking it is a step function, "jumping" (from 0) to value 1/20 at x = 16, retaining that value up to a further "jump" to 3/20 at x = 23, and so on. On the graph below, for convenience its values at x = 16, 23, …, 192 are shown, with these being joined by line segments.

$\hat{F}(x)$ is given by $1 - \hat{S}(x) = 1 - (1 + \hat{\lambda}x)e^{-\hat{\lambda}x}$ calculated at x = 16, 23, …, 192 using $\hat{\lambda}$ = 0.0283 as found in part (ii). The values of $\hat{F}(x)$ are given in the table. These also are plotted on the graph, joined by line segments for convenience.

x	16	23	35	44	51	55	67	97	127	192
$\hat{F}(x)$	0.076	0.139	0.261	0.354	0.423	0.461	0.565	0.759	0.874	0.972

Solution continued on next page



Key:

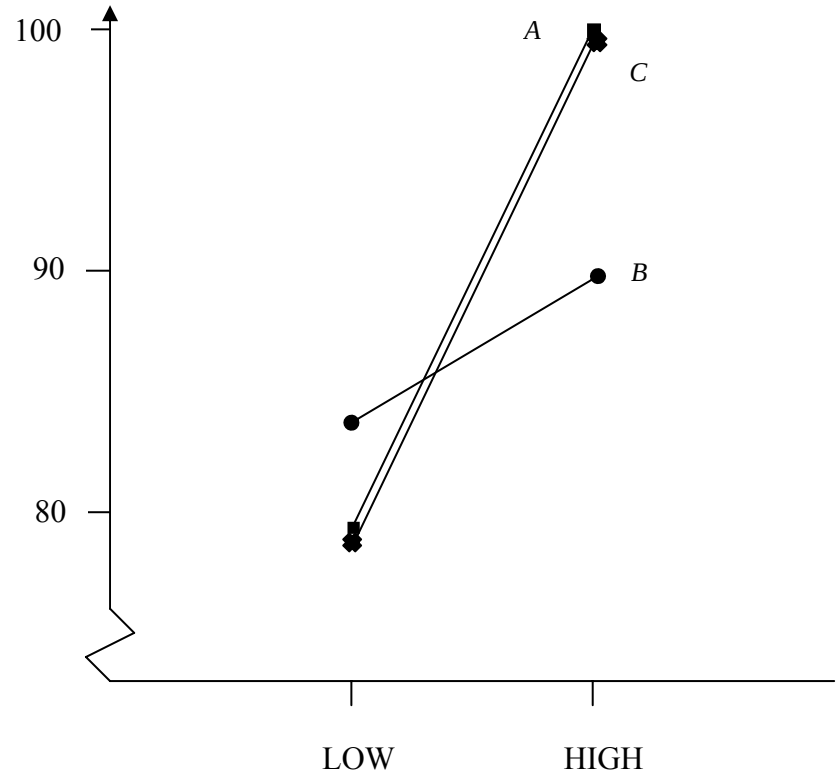
F^*

$\hat{F}$ (**NB** shown as just F on the graph)

F^* and $\hat{F}$ are close except between survival times of about 50 and 100, where $\hat{F}$ (the fitted model) somewhat underestimates the cumulative probability of dying – in that interval, patients are more likely to die than the fitted model predicts. In addition, the fitted model is slightly pessimistic towards the end of the range of survival times, so the result in part (iii) may be an underestimate of this probability of survival.

Higher Certificate, Paper III, 2006. Question 6

(i)



Interaction is when factors (here, the level and type of protein) do not appear to function independently. Here all the types – *A*, *B*, *C* – give an increase in mean weight gain with increasing level (although the behaviour of *B* is rather different from *A* and *C*). There is unlikely to be a large interaction – if there is any.

(ii) The analysis of variance table is as follows. Entries in *italics* are given in the question. The others need to be calculated.

SOURCE	DF	SS	MS	*F* value
Level	1	*3776.3*	3776.30	17.60 Compare $F_{1,54}$
Type	2	*82.5*	41.25	0.19 Compare $F_{2,54}$
Level * Type (Interaction)	2	*730.1*	365.05	1.70 Compare $F_{2,54}$
Error (Residual)	54	11586.0	214.56	$= \hat{\sigma}^2$
TOTAL	59	*16174.9*		

Solution continued on next page

Upper critical points of $F_{1,50}$ and $F_{2,50}$ are taken from the Society's statistical tables for use in examinations. Values for (1, 54) and (2, 54) will be very similar.

	5%	1%	0.1%
$F_{1,50}$	4.03	7.17	12.22
$F_{2,50}$	3.18	5.06	7.96

The *F* value for level is very highly significant; we have very strong evidence that the two levels of protein do not result in the same overall mean weight gain.

The *F* value for type is insignificant. We have no evidence to suggest that the three protein types are different in terms of the overall mean weight gain.

Similarly, we have no evidence that there is any interaction, i.e. that any protein type behaves differently as level is increased (even though the *B* responses are somewhat different from those of *A* and *C*). The graph in part (i) shows the pattern, and the analysis here confirms which are the significant sources of variation.

(iii) The 5 degrees of freedom for factors and interaction explain only 28.4% of the total variation (SS total). This is uncomfortably small.

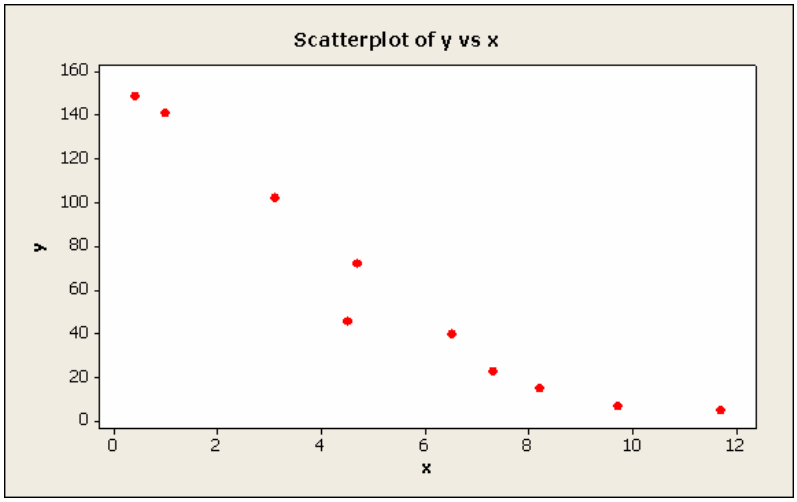
We note also that the estimate of experimental error is $\hat{\sigma}^2 = 214.56$ ($\hat{\sigma} = 14.65$), which is quite large compared with the values of the observations themselves (of order 100).

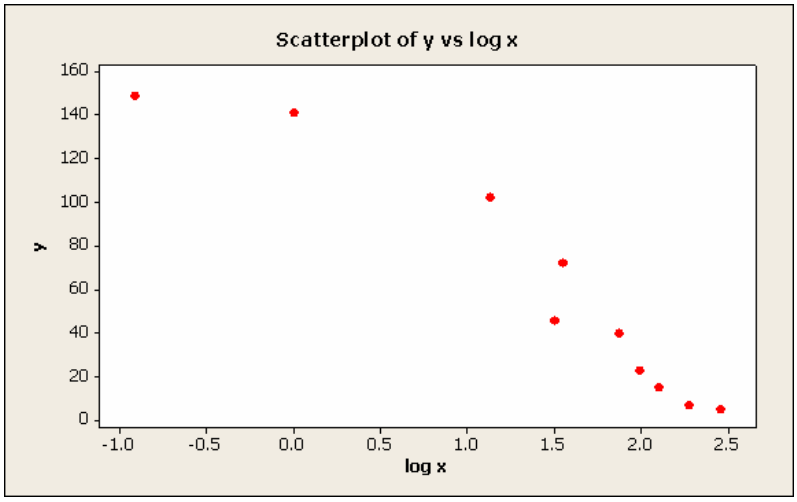
Perhaps it is simply the case that the weight gains are naturally very variable; or perhaps they are influenced by other covariates (e.g. initial weight).

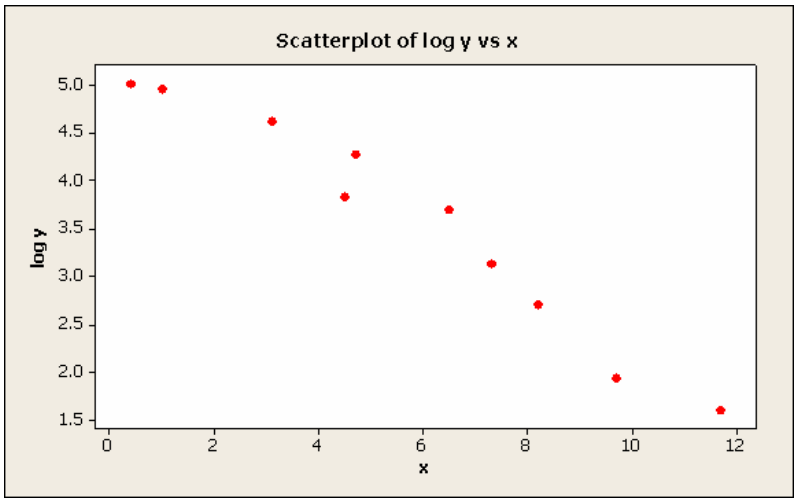
The dependence of weight gain on protein level appears strong and is intuitively appealing. But there may be more to learn about the response variable.

Higher Certificate, Paper III, 2006. Question 7

(i)







Solution continued on next page

There is some curvature in all these plots; $\log y$ on x is slightly "straighter" than y on x. Using $\log x$ looks worst. So use $\log y$ on x.

(ii) (c) has the highest R^2 – though all are good.

(c) also has the highest t values for the coefficients – though again all are good.

(c) also has the lowest residual variance relative to the mean response.

(c) is the only one without a "large" residual.

(c) appeared (marginally) the best plot.

It does not seem sensible to regress $\log y$ on $\log x$; it looks as if this would increase the curvature.

(iii) Using (c), we have $\log y = 5.41 - 0.322x$, so

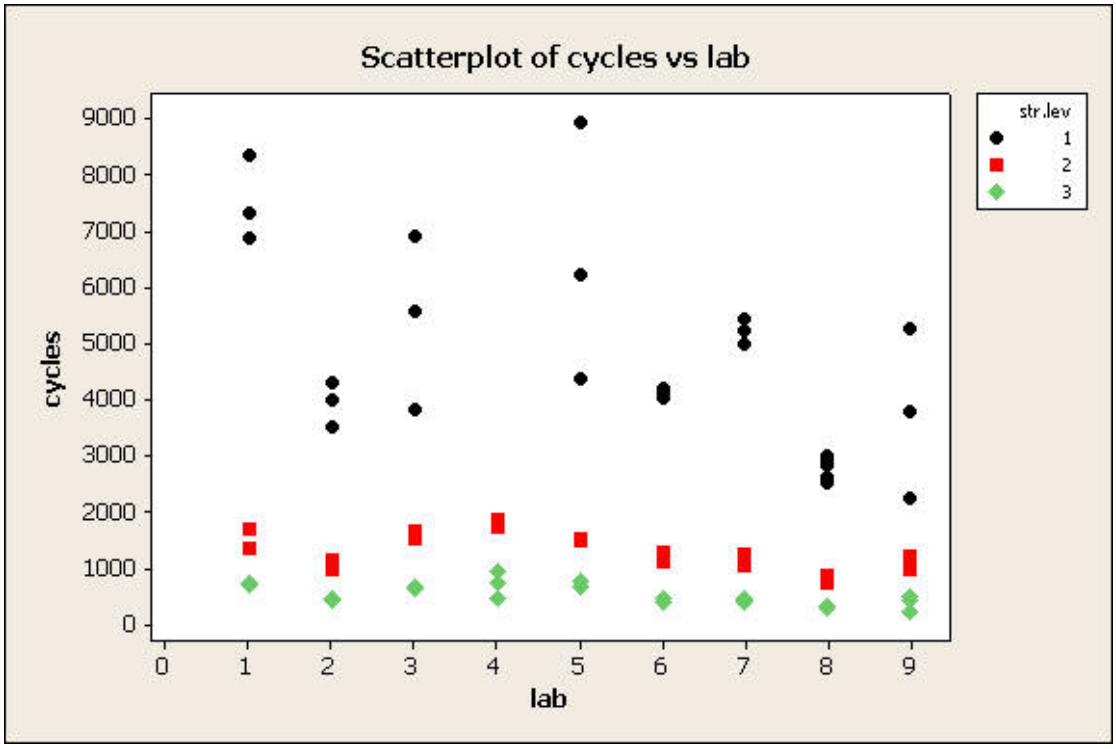
$$y = \exp\left(5.41 - 0.322x\right) = e^{5.41}e^{-0.322x} = 223.63e^{-0.322x}.$$

(iv) From the expression in part (iii), inserting $x = 5$, $y = 223.63e^{-1.61} = 44.7$. This is in hundreds per square kilometre. So the estimate is 4470 per square kilometre.

(v) Prediction within the range of the data may be adequate, except perhaps near the upper end because of the tendency for curvature there. Extrapolation to values of x outside the data will, for similar reasons, be unreliable, and the linear model is likely to underestimate density. Where is the next city or town centre? Interaction with that is very likely unless it is a long distance away. There may also be directional effects, i.e. densities changing more or less slowly according to the direction from the centre.

Higher Certificate, Paper III, 2006. Question 8

It is useful to have a graph showing the measurements made by each laboratory on each strain (laboratories A – J are relabelled 1 – 9). Individual responses are plotted.



The substantial difference in mean levels between the three strains makes it hard to present the graphical results on a convenient scale. However, several points emerge.

(1) Some laboratories have consistently lower readings than others. For example, H has by some way the lowest mean throughout; A, C, E are high; B, J tend to be low. Variability within laboratories is also very different; this can be seen especially for strain 1, where C, E, J give very wide ranges while F, G, H do not. The basic material used does not appear to have been so variable, because not all within-laboratory variation is large; technical reasons in respect of resources of equipment or people is a more likely reason.

(2) Level 1 is the lowest strain level, and it shows much higher means and much more variation than levels 2 and 3. There is clearly an inverse relationship between cycles to fatigue and strain level. However, a model assuming constant variance would not be suitable; transformations of the y (cycles) variable could be explored, possibly $\log y$. Prediction will be more accurate at higher strain levels, and should only be attempted within the range of levels already tested; extrapolation below level 1 or above level 3 would be unwise.

Overall, the laboratories lack consistency. If the aim is to have a laboratory-independent prediction, laboratory practice needs to be more consistent. If this cannot be achieved, the model used needs to incorporate a laboratory effect.